# Learning Explicit Contact for Implicit Reconstruction of Hand-held Objects from Monocular Images

**Junxing Hu[§†], Hongwen Zhang[‡], Zerui Chen[♮], Mengcheng Li[♯], Yunlong Wang[†], Yebin Liu[♯], Zhenan Sun[§†*]**

[§]University of Chinese Academy of Sciences, [†]MAIS CASIA, [‡]Beijing Normal University, [♮]Inria, [♯]Tsinghua University

## Abstract

Reconstructing hand-held objects from monocular RGB images is an appealing yet challenging task. In this task, contacts between hands and objects provide important cues for recovering the 3D geometry of the hand-held objects. Though recent works have employed implicit functions to achieve impressive progress, they ignore formulating contacts in their frameworks, which results in producing less realistic object meshes. In this work, we explore how to model contacts in an explicit way to benefit the implicit reconstruction of hand-held objects. Our method consists of two components: *explicit contact prediction* and *implicit shape reconstruction*. In the first part, we propose a new subtask of directly estimating 3D hand-object contacts from a single image. The part-level and vertex-level graph-based transformers are cascaded and jointly learned in a coarse-to-fine manner for more accurate contact probabilities. In the second part, we introduce a novel method to diffuse estimated contact states from the hand mesh surface to nearby 3D space and leverage diffused contact probabilities to construct the implicit neural representation for the manipulated object. Benefiting from estimating the interaction patterns between the hand and the object, our method can reconstruct more realistic object meshes, especially for object parts that are in contact with hands. Extensive experiments on challenging benchmarks show that the proposed method outperforms the current state of the arts by a great margin. Our code is publicly available at *https://junxinghu.github.io/projects/hoi.html*.

## Introduction

Reconstructing human-object interaction from monocular images is essential to understand the interactions between humans and the physical world. Toward this goal, recent progress has been achieved in the individual reconstruction of the body (Kocabas et al. 2021; Zhang et al. 2020, 2023), hands (Romero, Tzionas, and Black 2017; Kulon et al. 2020; Baek, Kim, and Kim 2019; Hampali et al. 2022; Chen et al. 2021; Boukhayma, Bem, and Torr 2019; Li et al. 2022), objects (Chen and Zhang 2019; Mescheder et al. 2019; Park et al. 2019; Groueix et al. 2018; Wang et al. 2018; Peng et al. 2021a), and their joint reconstruction (Hasson et al. 2019, 2020; Karunratanakul et al. 2020; Yang et al. 2021; Chen et al. 2022, 2023b; Ye, Gupta, and Tulsiani 2022). However, this task remains very challenging due to the complexity of hand poses and the diversity of interacting objects.

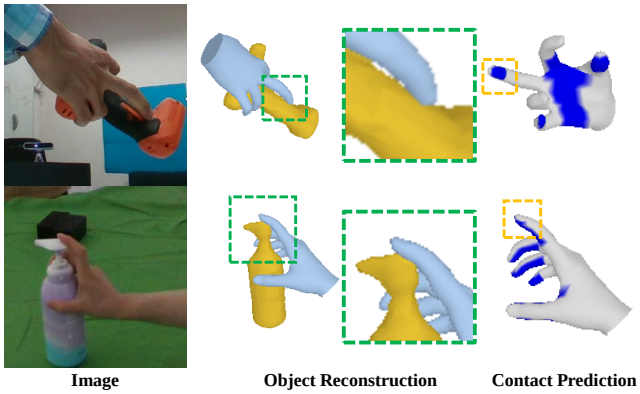



Figure 1: Given an RGB image, the proposed method predicts hand-object contacts and recovers the 3D geometry of the object. The insight is that the contacts could provide effective cues for the hand-held object reconstruction.

As hand-held objects involve the grasp configuration between hands and objects, the contacts play essential roles in modeling hand-object interactions. To improve the interaction, there have been several attempts to model the contact in different representations, including using contacts to optimize meshes (Hasson et al. 2019), the contact potential field (Yang et al. 2021), or the grasping field (Karunratanakul et al. 2020). However, these methods only model contacts as an additional loss function, which miss the chance to construct and exploit contact priors to simplify the 3D reconstruction problem.

Our key observation is that the contacts between hands and objects provide important cues for recovering the 3D geometry of the hand-held object. Modeling contacts between hands and objects can compensate for the lack of 3D information in monocular RGB images and makes it easier to infer the shape of the hand-held objects, especially for parts that are in contact with hands as shown in Fig. 1. Though previous methods have included contact losses (Hasson et al. 2019; Karunratanakul et al. 2020) or optimization objectives (Yang et al. 2021; Grady et al. 2021) in their reconstruction pipelines, they do not consider the usage of con-

*Corresponding author.

tacts as an intermediate representation to benefit the 3D reconstruction. In this work, we explore how to construct contact priors from the monocular RGB image to help recover the 3D object geometry. Specifically, we first predict contact points explicitly on the hand mesh surface. To our knowledge, estimating contact states from a single RGB image is explored only for human body mesh (Huang et al. 2022; Fieraru et al. 2020, 2021; Chen et al. 2023a) without focusing on the hand. To this end, we introduce a novel coarse-to-fine learning framework to jointly learn part-level and vertex-level contact states. In addition, we utilize the graph-based transformer which combines graph convolutions with transformers to better accumulate relevant features among adjacent nodes in the hand mesh and obtain more robust contact predictions.

Then, we attempt to exploit predicted contact states to simplify the 3D reconstruction task. Here, we follow the previous work (Ye, Gupta, and Tulsiani 2022) to model hand-held objects with deep implicit functions (Park et al. 2019), which can generate realistic and high-resolution object meshes. However, how to make implicit functions take good advantage of estimated contact states is also challenging and remains unsolved. The main challenge is that contact points are distributed on the hand surface in the discrete form, while implicit functions have continuous values in the whole 3D volume. To tackle the difficulty, we employ sparse convolutions to diffuse these discrete contact states from the hand surface to the 3D space. Then, the implicit function can naturally query corresponding contact features for a given 3D point and improve the neural implicit reconstruction. We conduct extensive experiments on HO3D (Hampali et al. 2020) and OakInk (Yang et al. 2022b) benchmarks to show that our method can reconstruct high-quality object meshes that interact faithfully with hands.

To sum up, the main contributions of this work can be listed as follows:

- We propose to leverage contact priors for better reconstruction of hand-held objects. To estimate contact states more accurately, we introduce a novel framework that jointly improves part-level and vertex-level contact states in a coarse-to-fine manner.
- To make discrete contact states compatible with continuous implicit shape functions, we propose to diffuse contact features from the hand mesh surface to the whole 3D volume, which enables the continuous query of contact features for implicit object reconstruction.
- We conduct extensive experiments on HO3D and OakInk benchmarks to validate the effectiveness of our method. Our method can produce more realistic hand-held object meshes and advance state-of-the-art accuracy.

## Related Work

Our work focuses on reconstructing hand-held objects from monocular RGB images. In this section, we first review related works in the field of 3D hand-object reconstruction. Then, we discuss how to leverage contact information to improve the quality of 3D reconstruction.

**3D Hand-object Reconstruction**. This task aims to reconstruct the 3D geometry of hands and hand-held objects from images. Existing approaches can be generally classified into two categories: multi-view and single-view methods. Multi-view methods (Hampali et al. 2020; Yang et al. 2022b; Chao et al. 2021; Oikonomidis, Kyriazis, and Argyros 2011; Wang et al. 2013) employ multiple cameras positioned at different viewpoints to infer the 3D structure of the grasping scenario. Though this type of method can generate very accurate 3D reconstruction results, they need careful camera calibrations and are inconvenient to deploy in the wild scene. Single-view methods only need monocular sensors (Ye, Gupta, and Tulsiani 2022; Hasson et al. 2019, 2020; Yang et al. 2021; Karunratanakul et al. 2020; Chen et al. 2022, 2023b; Tse et al. 2022a; Zhang et al. 2021; Hu et al. 2022; Kyriazis and Argyros 2014; Zhao et al. 2022) as inputs and are flexible to apply in real practice. In this work, we use the most common monocular RGB images as inputs. However, given the ill-posed nature, it is quite challenging to infer the 3D structure only from monocular RGB cues. To alleviate the difficulty of the hand reconstruction problem, Hasson *et al.* (Hasson et al. 2019, 2020) propose to employ the parametric hand model MANO (Romero, Tzionas, and Black 2017), which encodes rich hand priors, to predict the hand mesh. To produce more realistic hand meshes, recent works (Karunratanakul et al. 2020; Chen et al. 2022, 2023b) employ the neural implicit function (Park et al. 2019) to model the hand shape and use estimated hand pose priors (Chen et al. 2022, 2023b) to simplify the hand shape learning. However, compared with the hand part, hand-held object reconstruction is even more challenging. Since there are thousands of manipulated objects in our daily lives, it is difficult to make a unified object mesh template like MANO or estimate 6D poses reliably for diverse objects, especially for symmetric objects. Given its difficulty, some existing works (Yang et al. 2021; Hasson et al. 2020; Yang et al. 2022a) even make a strong assumption that the perfect object model is known at test time and only predicts its 6D pose. A recent work (Ye, Gupta, and Tulsiani 2022) relaxes this assumption and proposes to leverage estimated hand poses to benefit the model-free reconstruction of hand-held objects. In this work, we go a step further and argue that contacts between hands and objects could provide important cues for 3D reconstruction and introduce a novel framework to generate more realistic object meshes that interact with hands.

**Contacts in Object Reconstruction and Manipulation**. Learning to model and reconstruct the 3D geometry of objects from monocular images has been a crown jewel in the field of computer vision (Roberts 1963; Mundy 2006). Previous works usually represent the 3D object using explicit representations (*e.g.*, meshes (Groueix et al. 2018; Wang et al. 2018), point clouds (Qi et al. 2017a,b) or voxels (Choy et al. 2016; Riegler, Ulusoy, and Geiger 2017; Pavlakos et al. 2017)) and use deep neural networks to predict them. In recent years, neural implicit functions (Park et al. 2019; Mescheder et al. 2019; Chen and Zhang 2019) have gradually become a popular paradigm for 3D reconstruction. It is seamlessly compatible with neural networks and can theoretically reconstruct objects at unlimited resolution. How-

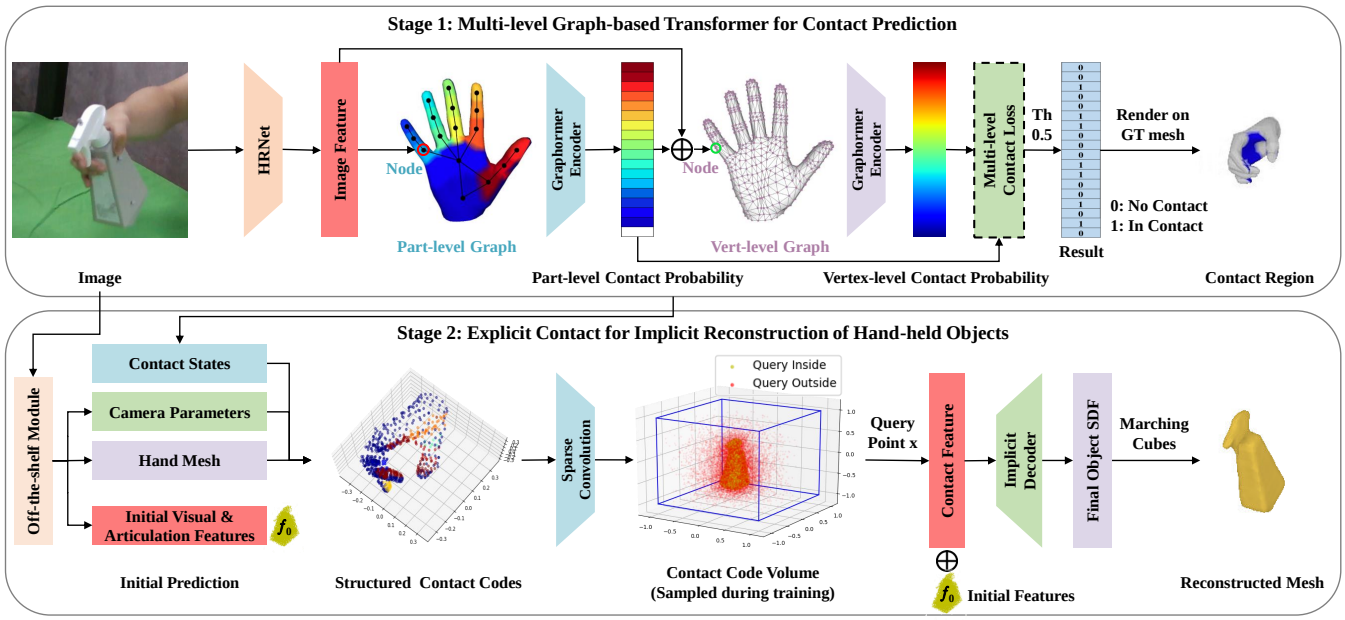



Figure 2: The overview of learning explicit contact for implicit reconstruction. First, the method estimates hand contact regions given a monocular RGB image. Based on the template hand mesh, part- and vertex-level graph-based transformers are cascaded for accurate predictions. Second, the estimated contact is used to construct the implicit neural representation. An off-the-shelf module is utilized to produce the camera parameters, hand mesh, and initial features. Then, the structured contact codes are generated by anchoring contact probabilities to the hand mesh surface. After sparse convolutions, the contact states on the hand surface are diffused to its nearby 3D space, which facilitates the perception and reconstruction of the manipulated object.

ever, the implicit function itself does not contain object surface priors, which makes it hard to fit diverse object surfaces. In this work, we construct a surface prior using contacts between hands and objects and simplify the learning problem. Actually, some works in robotics (Bicchi and Kumar 2000; Li et al. 2020; Buescher et al. 2015; Yin et al. 2023; Tse et al. 2022b) have shown that contacts could provide rich cues about the object shape and how to manipulate the given object. Some recent systems (Yin et al. 2023; Jain et al. 2019) can successfully manipulate different objects by using contact sensors. However, how to use contacts to benefit hand-held reconstruction is under-explored in our task. Previous works only use contact information in an implicit way. Methods using explicit object representations (Hasson et al. 2019; Yang et al. 2021; Grady et al. 2021) introduce contact loss terms to encourage objects to be close to reconstructed hand meshes. Some recent efforts (Karunratanakul et al. 2020; Ye, Gupta, and Tulsiani 2022) also introduce contact loss terms in the context of neural implicit representation. Different from them, we model and predict contact states explicitly and successfully leverage volume encoding (Peng et al. 2021b; Kwon et al. 2021; Choi et al. 2022) to diffuse contact information from hand surfaces to the 3D space for the hand-held object reconstruction.

## Method

In this section, we describe the technical details of the proposed method. As shown in Fig. 2, our method consists of two stages: explicit contact prediction and implicit shape reconstruction. In the first stage, we propose to predict part-level and vertex-level hand contact states in a coarse-to-fine manner. A graph-based transformer model is introduced to estimate contact probabilities more accurately. In the second stage, we present a novel method to leverage estimated contact states to improve the neural implicit reconstruction of hand-held objects.

### Explicit Contact Prediction

Given a single RGB image $I$, our method first predicts the contact regions between the hands and objects. Specifically, we estimate contact probabilities within $[0, 1]$ on hand meshes to measure the likelihood of the region touching the object. In our method, the contact probabilities are predicted from coarse to fine and denoted as $\boldsymbol{C_p} = \{\boldsymbol{c_p}^i \in [0,1]\}_{i=1}^{N_p}$ and $\boldsymbol{C_v} = \{\boldsymbol{c_v}^i \in [0,1]\}_{i=1}^{N_v}$ for the part-level and vertex-level contacts, where $N_p$ and $N_v$ are the number of the hand parts and hand mesh vertices, respectively.

**Multi-level Contact Graphs.** For more accurate predictions of the contact probabilities, multi-level contact graphs are leveraged to process the surface regions in the part and vertex levels such that the contact can be jointly learned from coarse to fine. Considering that the hand mesh can be naturally represented as a graph, we build the contact graphs based on the template MANO mesh (Romero, Tzionas, and Black 2017). Specifically, the part-level graph $G_p$ with $N_p$ nodes is generated relying on a coarse division of the hand regions. According to statistical contact frequency, the hand

surface is divided into $N_p$ subregions, including $(N_p - 1)$ subregions on the hand palm and one subregion on the back side of the hand. When building graph $G_p$, the center point of each part of the MANO template is taken as a graph node. For each graph node, its features are the concatenation of the image-based feature and its 3D coordinates. As shown in the first stage in Fig. 2, an image feature $f \in \mathbb{R}^D$ with the length of $D$ is extracted from $I$ by using an HRNet backbone (Wang et al. 2020). Therefore, each part-level graph node feature of $G_p$ is $\boldsymbol{g_p^i} \in \mathbb{R}^{D+3}, i = \{1, 2, \ldots, N_p\}$ and the adjacency matrix is encoded as the physical contact relationship between nodes. On the other hand, the vertex-level graph $G_v$ is generated based on the $N_v$ mesh vertices with an adjacency matrix from the MANO template. In addition to the image feature, the vertex-level node features of $G_v$ also include the part-level contact probability $\boldsymbol{C_p}$, resulting in the node feature $\boldsymbol{g_v^i} \in \mathbb{R}^{D+N_p+3}, i = \{1, 2, \ldots, N_v\}$.

**Graph-based Transformer for Contact Prediction.** In hand-object interaction, the contact area is usually occluded by hands or objects, which requires the network to perceive local details and global information. Following Graphormer (Lin, Wang, and Liu 2021), our contact estimators are designed as graph-based transformers that incorporate the graph convolution (Kipf and Welling 2017) into the transformer block (Vaswani et al. 2017). In this way, the graph convolution focuses on fine-grained local interactions, while the latter encodes the global relationships of the whole hand regions. As the contacts are predicted at the part and vertex levels, the architectures of the coarse and fine contact estimators are also built upon the graphs $G_p$ and $G_v$, respectively. Specifically, the coarse and fine contact estimators have $N_p$ and $N_v$ input tokens, which correspond to the same number of nodes in the graphs. Moreover, the two contact estimators have different hidden sizes in their transformer blocks. In practice, we find that a hidden size of 256 is sufficient for the part-level contact estimation, and the three blocks with hidden sizes of 1024, 256, and 64 work well for the vertex-level contact estimator.

For both the two contact estimators, the size of the output token is set to one. Similar to the settings in BSTRO (Huang et al. 2022), a sigmoid function is used to convert output tokens to contact probabilities in the range of [0, 1], and we extract contact points with probabilities greater than 0.5.

## Explicit Contact for Implicit Object Reconstruction

As shown in the second stage in Fig. 2, given the explicit contact prediction $\boldsymbol{C_p}$ and $\boldsymbol{C_v}$ with the hand mesh, our method first builds structured contact codes in a normalized 3D space. Then, they are fed into a sparse convolutional network to generate the contact code volumes $\boldsymbol{V}$ at different resolutions. This operation diffuses the contact states on the hand surface to the nearby 3D space and can be sampled continuously as additional conditions for the implicit reconstruction of objects.

**Initial Prediction.** Given an RGB image, an off-the-shelf module from IHOI (Ye, Gupta, and Tulsiani 2022) is used to generate the camera parameters, the hand mesh, and initial features $f_0$ including visual and articulation embeddings. By using the camera parameters, sampled 3D query points on the object surface are transformed into a normalized coordinate system around the hand wrist, which serve as the inputs for the subsequent structured contact codes.

**Structured Contact Codes.** The predicted contact states $\boldsymbol{C_v} \in \mathbb{R}^{N_v}$ are utilized to construct structured contact codes, which act as intermediate contact features. In the context of implicit reconstruction, we perform trilinear interpolation on estimated contact probabilities according to the contact point's position. In addition, to facilitate the network learning, each contact code $\boldsymbol{c_v^i} \in \mathbb{R}^1$ is mapped to a higher dimensional space by using the positional encoding (Mildenhall et al. 2022).

**Contact Code Volume.** There are two disadvantages of directly extracting features from structured contact codes. First, the contact information is only limited to the mesh surface and cannot cover the surrounding space of the hand where the object is located. Second, the vertices are too sparse in 3D space to provide enough contact information as most extracted features are zero vectors. Since the implicit functions have continuous values in the 3D volume, the sparse convolutions (Graham, Engelcke, and Van Der Maaten 2018) are utilized to diffuse the discrete contact states to the continuous space. Specifically, the structured contact codes are first scaled into the initial volume $V_0$ as the input. Then, a sparse convolutional network is used to process the contact code volumes $\boldsymbol{V} = \{V_i\}_{i=1}^{L}$ at $L$ different resolutions inspired by Neural Body (Peng et al. 2021b). As a result, the contact code volumes are not limited to contact states at the hand mesh surface and contain diffused contact features for nearby 3D space, which is compatible with the continuous implicit functions.

**Implicit Decoding.** Contact code volumes of different resolutions are first normalized to the same scale [-1, 1]. Then, the contact feature $fc_i$ is extracted by interpolation according to the query point $x$ from each contact code volume $V_i$. The final contact feature $fc$ is obtained as the concatenation of features extracted from volumes of different resolutions:

$$fc = \bigoplus(fc_1, fc_2, \cdots, fc_L) \tag{1}$$

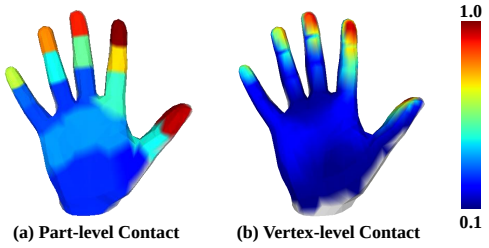
where $\bigoplus(\cdot)$ is a concatenation operation. After that, the SDF value $s$ on the query point $x$ can be computed via an implicit function $\mathcal{F}$ given the conditions of the contact feature $fc$ and the initial features $f_0$:

$$s = \mathcal{F}(x, fc, f_0) \tag{2}$$

Similar to other methods (Ye, Gupta, and Tulsiani 2022; Chen et al. 2022), the implicit function $\mathcal{F}$ is implemented as a decoder network similar to DeepSDF (Park et al. 2019), which composes of eight fully connected layers with a skip connection at the fourth layer.

## Training Details

**Contact Prediction.** In the first stage, the framework is trained in an end-to-end fashion to estimate the contact region from a single image. During training, the loss $\mathcal{L}_{Contact}$




(a) Part-level Contact (b) Vertex-level Contact

Figure 3: Visualization of contact frequency for different hand regions on OakInk (Yang et al. 2022b). (a) Part-level contact. (b) Vertex-level contact.

is used as follows:

$$\mathcal{L}_{Contact} = \lambda_p \mathcal{L}_{part} + \lambda_v \mathcal{L}_{vertex} + \lambda_{vs} \mathcal{L}_{vertex\_sub} \quad (3)$$

where $\lambda_p$, $\lambda_v$, and $\lambda_{vs}$ are balancing weights. $\mathcal{L}_{part}$, $\mathcal{L}_{vertex}$, and $\mathcal{L}_{vertex\_sub}$ are weighted binary cross entropy (BCE) losses between the ground truth and the predicted contact probabilities. The first one corresponds to the part-level contact. For multi-scale perception and computational efficiency, the template MANO mesh is downsampled to a sub-mesh with 195 vertices for graph generation. Processed by graphormer encoders, the coarse contact prediction is generated to compute the $\mathcal{L}_{vertex\_sub}$. Then, the coarse prediction is upsampled back to 778-dimensional refined results for $\mathcal{L}_{vertex}$. Fig. 3 illustrates the contact frequency on different hand regions by analyzing statistics on a large hand-object interaction dataset OakInk (Yang et al. 2022b). It can be observed that the frequencies of different regions vary greatly. Therefore, we normalize these frequencies to $[0.1, 1]$ and use them as weight priors to compute the weighted BCE losses.

**Implicit Reconstruction.** In the second stage, the off-the-shelf module from IHOI (Ye, Gupta, and Tulsiani 2022) is trained together with the proposed method. Similar to IHOI, the object model is provided to guide query point sampling during training, where 95% of the points are sampled around the model surface and others uniformly in the normalized space as shown in Fig. 2. It should be noted that at test time, query points are uniformly sampled in space since the object model is agnostic. In this part, the reconstruction loss $\mathcal{L}_{Recon}$ is calculated as follows:

$$\mathcal{L}_{Recon} = \mathcal{L}_{obj} + \mathcal{L}_{hoi} = \|s - \hat{s}\|_1 + \frac{1}{N_c} \sum_{i=1}^{N_c} (c_v^i \cdot |s_h^i|) \quad (4)$$

where $\mathcal{L}_{obj}$ is an L1 loss function between the ground truth $\hat{s}$ and the predicted SDF value $s$ of the object similar to other approaches (Ye, Gupta, and Tulsiani 2022; Chen et al. 2022, 2023b). $\mathcal{L}_{hoi}$ is related to $N_c$ vertices on the hand mesh that are in contact with the object (*i.e.*, the contact probability $c_v^i > 0.5$). $s_h^i$ is the SDF value calculated in Equation 2 of the hand contact vertices. Taking $c_v^i$ as the weight, $\mathcal{L}_{hoi}$ is the weighted average sum of the SDF values. This term serves as a regularization term to penalize hand contact points that penetrate or are far from the object.

# Experiments

## Implementation Details

In this work, the size of the hand-object centered image is $224 \times 224$. The number of graph nodes are $N_p = 18$ and $N_v = 778$. The length of the image feature is $D = 2048$. The shapes of contact code volumes ($L = 4$) are $V_0 = [64, 64, 64]$, $V_1 = [32, 32, 32]$, $V_2 = [16, 16, 16]$, $V_3 = [8, 8, 8]$, $V_4 = [4, 4, 4]$, and their code dimensions are $d_0 = 16$, $d_1 = 32$, $d_2 = 64$, $d_3 = d_4 = 128$. The balancing weights are $\lambda_p = 1$, $\lambda_v = \lambda_{vs} = 0.5$. The model is implemented by PyTorch (Paszke et al. 2019) and the HRNet backbone (Wang et al. 2020) is pre-trained on ImageNet (Wang et al. 2020). For both contact estimation and object reconstruction, the learning rate is set to 1e-4, and the Adam optimizer (Kingma and Ba 2015) is used. Each model is trained for 200 epochs on the RTX3090 GPU and the best results are reported.

## Datasets and Setup

The proposed method is evaluated on two challenging real-world datasets: OakInk (Yang et al. 2022b) and HO3D (Hampali, Sarkar, and Lepetit 2021). To our knowledge, they are two of the few benchmarks that provide official contact annotations and corresponding RGB images. OakInk is one of the latest and largest hand-object interaction datasets. It contains 230K images, capturing the single-hand interactions of 12 subjects with 100 objects from 32 categories. HO3D is a widely used dataset consisting of 103k images. The dataset captures 10 subjects interacting with 10 YCB objects (Calli et al. 2015). More detailed dataset settings are provided in the supplementary material.

## Evaluation Metrics

For contact prediction from a single image, standard detection metrics such as precision, recall, and F1-score are adopted. For the object reconstruction, the chamfer distance (CD, $mm$), F-score at 5mm and 10mm thresholds are reported. To evaluate the quality of the relation between objects and hands, the penetration depth (PD, $cm$) and intersection volume (IV, $cm^3$) are computed.

## Experimental Results for Contact Prediction

Since there is no specific method focused on predicting hand contact regions from monocular images, we first conduct ablation experiments on model settings, then compare and validate the effectiveness of the multi-level graphormer. Finally, we evaluate different levels of contact prediction.

**Ablation Study.** Table 1 illustrates the quantitative ablation results for vertex-level contact predictions on the OakInk dataset. $M_1$ is designed to estimate the hand mesh and vertex-level contact at the same time. It yields the overall lowest detection scores. Compared with $M_1$, $M_2$ further uses the loss $\mathcal{L}_{vertex\_sub}$ calculated on the sub-mesh proposed in Training Details. The precision, recall, and F1-score are improved by 5.6%, 11.4%, and 11.1%, respectively, proving the effectiveness of multi-scale features aggregation based on the hand model in this task. Different

Table 1: Ablation study for vertex-level contact predictions on the OakInk dataset. From left to right are whether to use $\mathcal{L}_{vertex\_sub}$ ($\mathcal{L}_{vs}$), whether to only predict contact (otherwise reconstruct the hand mesh at the same time), and whether to use the weighted loss.

| Method | $\mathcal{L}_{vs}$ | Only Contact | Weighted | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| $M_1$ | ✗ | ✗ | ✗ | 0.270 | 0.176 | 0.189 |
| $M_2$ | ✔ | ✗ | ✗ | 0.285 | 0.196 | 0.210 |
| $M_3$ | ✔ | ✔ | ✗ | 0.309 | 0.192 | 0.213 |
| $M_4$ | ✔ | ✔ | ✔ | **0.332** | **0.245** | **0.262** |

Table 2: Comparison of different network architectures on OakInk and HO3D benchmarks.

| Method | OakInk | | | HO3D | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| Single-Vertex | 0.332 | **0.245** | 0.262 | 0.476 | 0.422 | 0.416 |
| **Multi-Vertex** | **0.342** | 0.244 | **0.262** | **0.510** | **0.441** | **0.436** |
| Single-Part | 0.770 | 0.753 | 0.728 | 0.710 | 0.723 | 0.672 |
| **Multi-Part** | **0.790** | **0.767** | **0.747** | **0.722** | **0.741** | **0.685** |

from $M_2$, $M_3$ does not reconstruct the hand mesh and only performs hand contact prediction. Although the recall drops slightly, its precision improves by 8.4%, showing that focusing on a single task could make the network learn more effectively. Finally, compared with $M_3$, $M_4$ uses weight priors for BCE losses in Equation 2 and achieves a huge boost on all metrics (*e.g.*, 27.6% on recall and 23.0% on F1), showing that the weight priors of contacts introduced in Training Details can provide useful guidance for the model.

**Effectiveness of Multi-level Graphormer.** In this work, three network architectures are trained and evaluated on OakInk and HO3D benchmarks, respectively. As shown in Table 2, in addition to the multi-level graphormer encoders, we also use the single-level model in Fig. 2 for contact prediction. The outputs of the multi-level method are compared with corresponding single-level outputs. Though we observe that a single vertex-level model yields slightly better recall on OakInk, the multi-level one can improve the precision from 0.332 to 0.342 benefiting from using the part-level output to refine features for vertices. Regarding the part-level output, the coarse-to-fine model outperforms the single part-level model on all three evaluation metrics. The multi-level model also achieves superior performance on HO3D for all three metrics, which further demonstrates the advantage of using the proposed coarse-to-fine learning framework. Fig. 4 further illustrates the qualitative results of the proposed method on two benchmarks. Benefiting from accumulating both global contexts and local details by using the graph-based transformer, the proposed method is robust to input images with hand or object occlusions.

**Part-level vs. Vertex-level Prediction.** In Table 2, the vertex-level predictions are worse than the part-level results, showing that the dense vertex-level prediction is more difficult than the sparse one. For the single-level architecture,

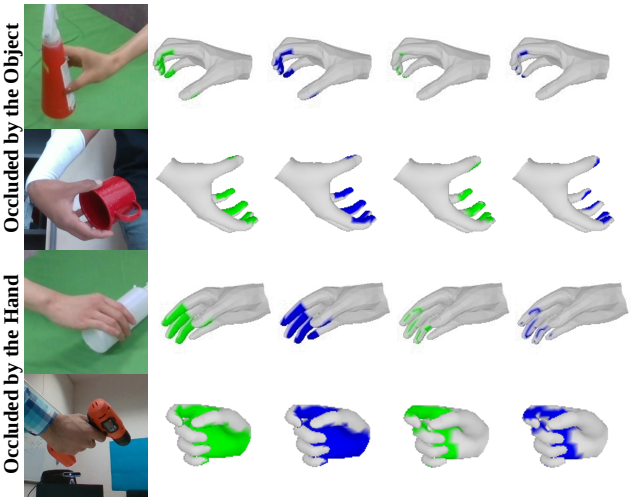



Figure 4: Visualizations of contact prediction on OakInk (Rows 1, 3) and HO3D (Rows 2, 4) datasets. Since the method only estimates contact, the result is rendered on the ground truth hand mesh. For samples whose contact regions are occluded by hands, hand meshes are rotated 180 degrees for clear visualization. The proposed method is robust to both hand and object occlusions.

Table 3: Ablation study for contact modeling in object reconstruction on HO3D. From left to right are whether to use the sparse convolution (SPC, otherwise nearest neighbor diffusion), whether to use the estimated contact (ESC, otherwise an all-one contact vector), and whether to use multi-scale contact code volumes (MSV, otherwise only $V_3 + V_4$).

| Method | SPC | ESC | MSV | F@5mm↑ | F@10mm↑ | CD $(mm)$↓ |
|---|---|---|---|---|---|---|
| $N_1$ | ✗ | ✔ | ✔ | 0.261 | 0.475 | 1.110 |
| $N_2$ | ✔ | ✗ | ✔ | 0.361 | 0.592 | 0.848 |
| $N_3$ | ✔ | ✔ | ✗ | 0.371 | 0.614 | 0.680 |
| $N_4$ | ✔ | ✔ | ✔ | **0.393** | **0.633** | **0.646** |

the F1 score of the vertex-level method on OakInk is only 0.262, while the single part-level model achieves 0.728. On the HO3D dataset, we can observe a similar performance gap between the part-level and vertex-level accuracy. Fig. 4 illustrates that part-level predictions are closer to the ground truth than vertex-level predictions. Therefore, part-level predictions are converted to vertex-level ones according to the fixed correspondence and then propagated to hand mesh vertices for subsequent experiments. More details, comparisons, and verification that contact prediction is positively correlated with object reconstruction are provided in the supplementary material.

## Experimental Results on Object Reconstruction

**Ablation Study.** Table 3 illustrates the ablations for contact modeling in object reconstruction on HO3D and $N_4$ is our final method. The ground truth hand meshes are adopted to ignore the influence of the hand pose. $N_1$ is designed

Table 4: Comparison with state-of-the-art methods on HO3D. '*' denotes using the ground truth hand mesh.

| Method | F@5mm | F@10mm | CD | PD $(cm)\downarrow$ | IV $(cm^3)\downarrow$ |
|---|---|---|---|---|---|
| HO | 0.110 | 0.220 | 4.190 | - | - |
| GF | 0.120 | 0.240 | 4.960 | - | - |
| IHOI | 0.280 | 0.500 | 1.530 | - | - |
| **Ours** | **0.313** | **0.542** | **1.081** | **1.02** | **5.11** |
| IHOI* | 0.351 | 0.600 | 0.656 | 0.90 | 4.10 |
| **Ours*** | **0.393** | **0.633** | **0.646** | **0.67** | **2.91** |

Table 5: Comparison on the OakInk benchmark.

| Method | F@5mm | F@10mm | CD | PD $(cm)$ | IV $(cm^3)$ |
|---|---|---|---|---|---|
| IHOI | 0.432 | 0.658 | 0.491 | 0.75 | 4.36 |
| Ours w/o $\mathcal{L}_{hoi}$ | 0.447 | 0.716 | 0.274 | 0.66 | 3.03 |
| **Ours** | **0.459** | **0.718** | **0.260** | **0.62** | **2.67** |

to diffuse the contact information by using a simple nearest neighbor method like LoopReg (Bhatnagar et al. 2020). It is slower and much worse than $N_4$ since the contact information is not fully learned like sparse convolution, which may even bring negative effects. $N_3$ only combines two contact code volumes (i.e., $V_3 + V_4$) and performs worse than $N_4$ with four scales, indicating the effectiveness of the full multi-scale contact code volumes. In addition, $N_2$ removes the contact estimation module and takes an all-one contact vector for a fair comparison. Its results (e.g., F@5mm $=$ 0.361) are worse than that of $N_4$ (e.g., F@5mm $=$ 0.393), showing that the estimated contacts could provide flexible and efficient guidance for object reconstruction. More ablation results can be found in the supplementary material.

**Quantitative Comparison on HO3D.** Since most prior methods require the object template during inference, the methods most relevant to ours are HO (Hasson et al. 2019), GF (Karunratanakul et al. 2020), and IHOI (Ye, Gupta, and Tulsiani 2022). For a fair comparison, we use the same predicted hands from (Rong, Shiratori, and Joo 2020) as IHOI and the estimated contact states from our multi-level model. As shown in Table 4, when our model uses predicted hand meshes, we observe that our method can improve F@5mm and F@10mm by 11.8% and 8.4% and greatly reduce the chamfer distance by 29.3%. When the hand mesh is perfect, our method also shows an obvious advantage and consistently outperforms IHOI across all metrics. It largely improves F@5mm by 12.0% and reduces the intersection volume by 29.0%, demonstrating the superiority of our method in model-free object reconstruction.

**Quantitative Comparison on OakInk.** To show that our model can work well for unseen objects, we split the dataset to make sure that testing objects do not exist in the training set. As shown in Table 5, when our model is not trained together with $L_{hoi}$, it can still outperform IHOI on all metrics. Our final model, which is learned with $L_{hoi}$, achieves even better results on different metrics. Compared with IHOI, our method can largely improve F@5mm and F@10mm by 6.3% and 9.1%, respectively. At the same time, it reduces the penetration depth and intersection volume by 17.3% and 38.8%, which suggests our model can reconstruct more realistic objects that naturally interact with hands.

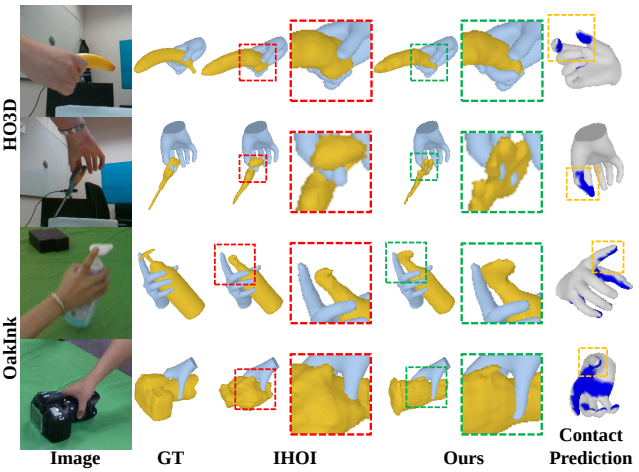



Figure 5: Qualitative comparison with the state-of-the-art method on the HO3D and OakInk datasets. Our method can reconstruct more realistic objects, especially for parts that are in contact with hands.

**Qualitative Comparison.** Fig. 5 illustrates qualitative comparisons on the HO3D and OakInk datasets. Compared with the state-of-the-art IHOI (Ye, Gupta, and Tulsiani 2022) (red dotted box), our method shows a clear advantage in the reconstruction of object parts that are in contact with the hand. It can be seen that the predicted hand contacts (yellow dashed box) provide effective guidance to recover corresponding object parts (green dashed box). We also observe that our method is robust to occlusions. As illustrated in the first, second, and fourth rows in Fig. 5, our model can still work well when objects are occluded by hands. For unseen objects with complex structures (*e.g.*, camera) in OakInk, our model can also obtain realistic results. More qualitative results can be found in the supplementary material.

## Conclusion

This paper introduces a novel representation of explicit contacts for the implicit reconstruction of hand-held objects. First, the multi-level graph-based transformer encoders are cascaded to estimate accurate 3D hand-object contacts from a single RGB image. Then, the predicted contact states are anchored to the hand surface and diffused to the nearby space to construct the implicit neural representation for the manipulated object. Extensive experiments on HO3D and OakInk datasets indicate that our method can pay more attention to the object parts that are in contact with hands and reconstruct more realistic object meshes. The proposed method currently focuses on hand-held object reconstruction. In future work, we attempt to integrate the hand reconstruction module for better hand-object interaction reconstruction and leverage object category priors to improve generalization.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. U23B2054, 62006225, 62071468, and also funded by the National Key Research and Development Program of China under Grant No. 2021ZD0113503 and 2022YFC3310400, the Fundamental Research Funds for the Central Universities under Grant No. 2233100028. We would like to thank Liang An and Yuxiang Zhang for their help, feedback, and discussions in the early work of this paper.

## References

Baek, S.; Kim, K. I.; and Kim, T.-K. 2019. Pushing the envelope for RGB-based dense 3D hand pose estimation via neural rendering. In *CVPR*.

Bhatnagar, B. L.; Sminchisescu, C.; Theobalt, C.; and Pons-Moll, G. 2020. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. *NeurIPS*.

Bicchi, A.; and Kumar, V. 2000. Robotic grasping and contact: A review. In *ICRA*.

Boukhayma, A.; Bem, R. d.; and Torr, P. H. 2019. 3D hand shape and pose from images in the wild. In *CVPR*.

Buescher, G.; Meier, M.; Walck, G.; Haschke, R.; and Ritter, H. J. 2015. Augmenting curved robot surfaces with soft tactile skin. In *IROS*.

Calli, B.; Singh, A.; Walsman, A.; Srinivasa, S.; Abbeel, P.; and Dollar, A. M. 2015. The ycb object and model set: Towards common benchmarks for manipulation research. In *ICAR*.

Chao, Y.-W.; Yang, W.; Xiang, Y.; Molchanov, P.; Handa, A.; Tremblay, J.; Narang, Y. S.; Van Wyk, K.; Iqbal, U.; Birchfield, S.; et al. 2021. DexYCB: A benchmark for capturing hand grasping of objects. In *CVPR*.

Chen, X.; Liu, Y.; Ma, C.; Chang, J.; Wang, H.; Chen, T.; Guo, X.; Wan, P.; and Zheng, W. 2021. Camera-space hand mesh recovery via semantic aggregation and adaptive 2D-1D registration. In *CVPR*.

Chen, Y.; Dwivedi, S. K.; Black, M. J.; and Tzionas, D. 2023a. Detecting Human-Object Contact in Images. In *CVPR*.

Chen, Z.; Chen, S.; Schmid, C.; and Laptev, I. 2023b. gSDF: Geometry-Driven Signed Distance Functions for 3D Hand-Object Reconstruction. In *CVPR*.

Chen, Z.; Hasson, Y.; Schmid, C.; and Laptev, I. 2022. AlignSDF: Pose-Aligned Signed Distance Fields for Hand-Object Reconstruction. In *ECCV*.

Chen, Z.; and Zhang, H. 2019. Learning implicit fields for generative shape modeling. In *CVPR*.

Choi, H.; Moon, G.; Armando, M.; Leroy, V.; Lee, K. M.; and Rogez, G. 2022. Mononhr: Monocular neural human renderer. In *3DV*.

Choy, C. B.; Xu, D.; Gwak, J.; Chen, K.; and Savarese, S. 2016. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *ECCV*.

Fieraru, M.; Zanfir, M.; Oneata, E.; Popa, A.-I.; Olaru, V.; and Sminchisescu, C. 2020. Three-dimensional reconstruction of human interactions. In *CVPR*.

Fieraru, M.; Zanfir, M.; Oneata, E.; Popa, A.-I.; Olaru, V.; and Sminchisescu, C. 2021. Learning complex 3D human self-contact. In *AAAI*.

Grady, P.; Tang, C.; Twigg, C. D.; Vo, M.; Brahmbhatt, S.; and Kemp, C. C. 2021. ContactOpt: Optimizing contact to improve grasps. In *CVPR*.

Graham, B.; Engelcke, M.; and Van Der Maaten, L. 2018. 3D semantic segmentation with submanifold sparse convolutional networks. In *CVPR*.

Groueix, T.; Fisher, M.; Kim, V. G.; Russell, B. C.; and Aubry, M. 2018. A papier-mâché approach to learning 3d surface generation. In *CVPR*.

Hampali, S.; Rad, M.; Oberweger, M.; and Lepetit, V. 2020. HOnnotate: A method for 3D annotation of hand and object poses. In *CVPR*.

Hampali, S.; Sarkar, S. D.; and Lepetit, V. 2021. HO-3D_v3: Improving the accuracy of hand-object annotations of the HO-3D dataset. *arXiv preprint arXiv:2107.00887*.

Hampali, S.; Sarkar, S. D.; Rad, M.; and Lepetit, V. 2022. Keypoint Transformer: Solving Joint Identification in Challenging Hands and Object Interactions for Accurate 3D Pose Estimation. In *CVPR*.

Hasson, Y.; Tekin, B.; Bogo, F.; Laptev, I.; Pollefeys, M.; and Schmid, C. 2020. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*.

Hasson, Y.; Varol, G.; Tzionas, D.; Kalevatykh, I.; Black, M. J.; Laptev, I.; and Schmid, C. 2019. Learning joint reconstruction of hands and manipulated objects. In *CVPR*.

Hu, H.; Yi, X.; Zhang, H.; Yong, J.-H.; and Xu, F. 2022. Physical Interaction: Reconstructing Hand-object Interactions with Physics. In *SIGGRAPH Asia*.

Huang, C.-H. P.; Yi, H.; Höschle, M.; Safroshkin, M.; Alexiadis, T.; Polikovsky, S.; Scharstein, D.; and Black, M. J. 2022. Capturing and inferring dense full-body human-scene contact. In *CVPR*.

Jain, D.; Li, A.; Singhal, S.; Rajeswaran, A.; Kumar, V.; and Todorov, E. 2019. Learning deep visuomotor policies for dexterous hand manipulation. In *ICRA*.

Karunratanakul, K.; Yang, J.; Zhang, Y.; Black, M. J.; Muandet, K.; and Tang, S. 2020. Grasping Field: Learning implicit representations for human grasps. In *3DV*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Kocabas, M.; Huang, C.-H. P.; Hilliges, O.; and Black, M. J. 2021. PARE: Part attention regressor for 3D human body estimation. In *ICCV*.

Kulon, D.; Güler, R. A.; Kokkinos, I.; Bronstein, M.; and Zafeiriou, S. 2020. Weakly-Supervised Mesh-Convolutional Hand Reconstruction in the Wild. In *CVPR*.

Kwon, Y.; Kim, D.; Ceylan, D.; and Fuchs, H. 2021. Neural human performer: Learning generalizable radiance fields for human performance rendering. *NeurIPS*.

Kyriazis, N.; and Argyros, A. 2014. Scalable 3D tracking of multiple interacting objects. In *CVPR*.

Li, M.; An, L.; Zhang, H.; Wu, L.; Chen, F.; Yu, T.; and Liu, Y. 2022. Interacting attention graph for single image two-hand reconstruction. In *CVPR*.

Li, Q.; Kroemer, O.; Su, Z.; Veiga, F. F.; Kaboli, M.; and Ritter, H. J. 2020. A review of tactile information: Perception and action through touch. *TRO*.

Lin, K.; Wang, L.; and Liu, Z. 2021. Mesh graphormer. In *CVPR*.

Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; and Geiger, A. 2019. Occupancy Networks: Learning 3D reconstruction in function space. In *CVPR*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2022. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*.

Mundy, J. L. 2006. Object Recognition in the Geometric Era: A Retrospective. In *Toward Category-Level Object Recognition*, Lecture Notes in Computer Science.

Oikonomidis, I.; Kyriazis, N.; and Argyros, A. A. 2011. Full DOF tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *ICCV*.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*.

Pavlakos, G.; Zhou, X.; Derpanis, K. G.; and Daniilidis, K. 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*.

Peng, S.; Jiang, C.; Liao, Y.; Niemeyer, M.; Pollefeys, M.; and Geiger, A. 2021a. Shape as points: A differentiable poisson solver. *NeurIPS*.

Peng, S.; Zhang, Y.; Xu, Y.; Wang, Q.; Shuai, Q.; Bao, H.; and Zhou, X. 2021b. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. PointNet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. PointNet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*.

Riegler, G.; Ulusoy, A. O.; and Geiger, A. 2017. OctNet: Learning Deep 3D Representations at High Resolutions. In *CVPR*.

Roberts, L. G. 1963. *Machine perception of three-dimensional solids*. Ph.D. thesis, Massachusetts Institute of Technology.

Romero, J.; Tzionas, D.; and Black, M. J. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *TOG*.

Rong, Y.; Shiratori, T.; and Joo, H. 2020. FrankMocap: Fast monocular 3D hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*.

Tse, T. H. E.; Kim, K. I.; Leonardis, A.; and Chang, H. J. 2022a. Collaborative Learning for Hand and Object Reconstruction with Attention-guided Graph Convolution. In *CVPR*.

Tse, T. H. E.; Zhang, Z.; Kim, K. I.; Leonardis, A.; Zheng, F.; and Chang, H. J. 2022b. S2Contact: Graph-Based Network for 3D Hand-Object Contact Estimation with Semi-supervised Learning. In *ECCV*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS*.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *TPAMI*.

Wang, N.; Zhang, Y.; Li, Z.; Fu, Y.; Liu, W.; and Jiang, Y.-G. 2018. Pixel2mesh: Generating 3D mesh models from single RGB images. In *ECCV*.

Wang, Y.; Min, J.; Zhang, J.; Liu, Y.; Xu, F.; Dai, Q.; and Chai, J. 2013. Video-based hand manipulation capture through composite motion control. *TOG*.

Yang, L.; Li, K.; Zhan, X.; Lv, J.; Xu, W.; Li, J.; and Lu, C. 2022a. ArtiBoost: Boosting Articulated 3D Hand-Object Pose Estimation via Online Exploration and Synthesis. In *CVPR*.

Yang, L.; Li, K.; Zhan, X.; Wu, F.; Xu, A.; Liu, L.; and Lu, C. 2022b. OakInk: A Large-scale Knowledge Repository for Understanding Hand-Object Interaction. In *CVPR*.

Yang, L.; Zhan, X.; Li, K.; Xu, W.; Li, J.; and Lu, C. 2021. CPF: Learning a contact potential field to model the hand-object interaction. In *ICCV*.

Ye, Y.; Gupta, A.; and Tulsiani, S. 2022. What's in your hands? 3D Reconstruction of Generic Objects in Hands. In *CVPR*.

Yin, Z.-H.; Huang, B.; Qin, Y.; Chen, Q.; and Wang, X. 2023. Rotating without Seeing: Towards In-hand Dexterity through Touch. *RSS*.

Zhang, H.; Cao, J.; Lu, G.; Ouyang, W.; and Sun, Z. 2020. Learning 3d human shape and pose from dense body parts. *TPAMI*, 44(5): 2610–2627.

Zhang, H.; Tian, Y.; Zhang, Y.; Li, M.; An, L.; Sun, Z.; and Liu, Y. 2023. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *TPAMI*, 45(10): 12287–12303.

Zhang, H.; Zhou, Y.; Tian, Y.; Yong, J.-H.; and Xu, F. 2021. Single depth view based real-time reconstruction of hand-object interactions. *TOG*.

Zhao, Z.; Zuo, B.; Xie, W.; and Wang, Y. 2022. Stability-driven contact reconstruction from monocular color images. In *CVPR*.